DATA Report-Drug-related Crime Analysis

Md Saddam Hosen
Matriculation number:23375480
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

1.The central question for this project is:

How can we integrate and analyze crime and arrest data to identify trends and correlations for improving law enforcement strategies in Los Angeles?

2. Data Sources

Chosen Datasets

Crime Data (2020–Present):

- **Source**: Los Angeles Open Data Portal (Link)
- **Content**: Information about crimes in Los Angeles, including incident IDs, locations, dates, and crime descriptions.
- **Reason for Choice**: This dataset provides a comprehensive record of crimes, which is essential for understanding crime trends.

Arrest Data (2020–Present):

- **Source**: Los Angeles Open Data Portal (Link)
- **Content**: Details about arrests, including incident IDs, dates, and charges.
- **Reason for Choice**: This dataset complements the crime dataset by linking crimes to arrests, enabling deeper correlation analysis.

Data Structure and Quality

- Both datasets are in CSV format and contain structured tabular data.
- Key columns include `Incident_ID` for linking, `Date` for temporal analysis, and `Latitude/Longitude` for geospatial analysis.
- Data quality issues encountered:
  - Missing values in critical columns (e.g., `Latitude`, `Longitude`).
  - Inconsistent date formats.
  - Overlapping and redundant data due to lack of normalization.

Licenses

- Both datasets are covered under the **City of Los Angeles Open Data License**, which allows free public use.
  - Obligations: Cite the source of the data and avoid re-identification of individuals.
  - Fulfillment: All citations are provided, and data is anonymized.

3. Data Pipeline

Overview: The data pipeline automates the process of downloading, cleaning, integrating, and analyzing the two datasets. It is implemented using Python with the following technologies:

- **Pandas**: Data cleaning and transformation.

- **Folium**: Visualization of geospatial data.
- **Matplotlib**: Plotting trends and analysis.

Pipeline Stages

1. **Data Loading**:
   a. Data is downloaded programmatically from the provided links.
   b. Columns are standardized for consistency (e.g., renaming `DATE OCC` to `Date`).
2. **Data Cleaning**:
   a. Converted `Date` columns to datetime format for analysis.
   b. Removed rows with missing or invalid data in critical columns (e.g., `Latitude`, `Longitude`).
3. **Data Integration**:
   a. Merged crime and arrest datasets on the `Incident_ID` column.
   b. Handled duplicate entries and ensured only 2020+ data was included.
4. **Error Handling**:
   a. Introduced checks for missing columns and invalid data types.
   b. If critical columns (`Date`, `Incident_ID`) are missing, the pipeline raises errors and halts.

Challenges and Solutions

- **Challenge**: Missing geospatial data (latitude/longitude).
  - **Solution**: Filtered out rows with missing values and flagged data gaps for future review.
- **Challenge**: Different date formats in the datasets.
  - **Solution**: Used `pd.to_datetime()` with `errors='coerce'` to standardize formats.

Meta-Quality Measures

- Logs all errors encountered during data loading and transformation.
- Automatically adjusts for changing column names by using configurable mappings.

4. Results and Limitations

Output Data

- The final dataset combines crime and arrest data with the following key columns:
  - **Incident_ID**: Unique identifier linking crimes to arrests.

    - **Date**: Standardized date format for trend analysis.
    - **Latitude/Longitude**: Geospatial data for visualization.
    - **Crime Code Description**: Crime details.

Data Structure and Quality

- The output is a clean CSV file with 50,000+ rows and ~10 columns.
- Missing values in geospatial data result in slight underrepresentation in geospatial visualizations.

Output Format

- **Format**: CSV (Comma-Separated Values).
    - **Reason**: Widely compatible format for further analysis and visualization tools.

Limitations

1. **Data Gaps**:
    a. Missing location data limits the scope of geospatial analysis.
    b. Arrest data may not represent all crimes due to underreporting or unlinked cases.
2. **Timeliness**:
    a. Data is updated periodically, and the pipeline does not yet account for real-time updates.

5. Conclusion

This project successfully integrates two key datasets to analyze crime and arrest trends in Los Angeles. The automated pipeline ensures efficient processing, cleaning, and integration of the data while handling errors gracefully. Future improvements could include addressing geospatial data gaps and incorporating real-time updates for more dynamic analysis.